

AD-753 463

A SURVEY OF ERROR ESTIMATES FOR ITERATIVE  
SOLUTIONS OF SYSTEMS OF LINEAR EQUATIONS  
WITH AN APPLICATION TO THE SOLUTION OF  
POISSON'S EQUATION

Michael J. Vander Vorst

Naval Ordnance Laboratory  
White Oak, Maryland

3 October 1972

DISTRIBUTED BY:

**NTIS**

National Technical Information Service  
U. S. DEPARTMENT OF COMMERCE  
5285 Port Royal Road, Springfield, Va. 22151

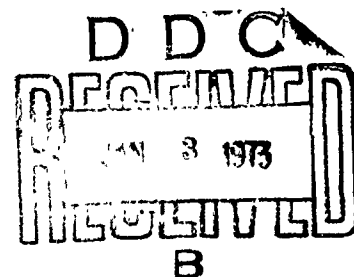
AD753463

NOLTR 72-189

A SURVEY OF ERROR ESTIMATES FOR ITERATIVE  
SOLUTIONS OF SYSTEMS OF LINEAR EQUATIONS  
WITH AN APPLICATION TO THE SOLUTION OF  
POISSON'S EQUATION

3y  
Michael J. Vander Vorst

3 OCTOBER 1972



NOL

NAVAL ORDNANCE LABORATORY, WHITE OAK, SILVER SPRING, MARYLAND

NOLTR 72-189

NATIONAL TECHNICAL  
INFORMATION SERVICE

APPROVED FOR PUBLIC RELEASE;  
DISTRIBUTION UNLIMITED

Security Classification

## DOCUMENT CONTROL DATA - R &amp; D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified

1. ORIGINATING ACTIVITY (Corporate author) Naval Ordnance Laboratory White Oak, Silver Spring, Maryland 20910		2a. REPORT SECURITY CLASSIFICATION Unclassified	
		2b. GROUP	
3. REPORT TITLE A Survey of Error Estimates for Iterative Solutions of Systems of Linear Equations with an Application to the Solution of Poisson's Equation.			
4. DESCRIPTIVE NOTES (Type of report and inclusive dates)			
5. AUTHOR(S) (First name, middle initial, last name) Michael J. Vander Vorst			
6. REPORT DATE 3 October 1972		7a. TOTAL NO. OF PAGES 35 4/	7b. NO. OF REFS 21
8a. CONTRACT OR GRANT NO.		9a. ORIGINATOR'S REPORT NUMBER(S) NOLTR No. 72-189	
b. PROJECT NO. MAT-03L-000/ZR00-001-010.		9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	
c.			
d.			
10. DISTRIBUTION STATEMENT Approved for public release; distribution unlimited.			
11. SUPPLEMENTARY NOTES		12. SPONSORING MILITARY ACTIVITY Chief of Naval Material (NAV MAT (03)) Department of the Navy Washington, D. C. 20390	
13. ABSTRACT In the solution of a set of linear equations $Ax = b$ by an iterative method one would like an estimate of the error $x - x_n$ so that the iteration can be stopped after the error is within acceptable bounds. We present a review of several types of error estimates paying particular attention to estimates of the form $\ x_{n+1} - x\  \leq c \ x_{n+1} - x_n\ $ for the class of iterative methods of the form $x_{n+1} = Mx_n + k$ . In addition we give a new error estimate for the successive over-relaxation method which is a generalization of an estimate of Sassenfeld for the Gauss-Seidel method. We present a numerical example of an error estimate of Albrecht for the successive over-relaxation method applied to the iterative solution of the system of linear equations which arise from the finite difference formulation of the Poisson equation.			

1A

DD FORM 1473 (PAGE 1)

1 NOV 65

S/N 0101-807-6801

Security Classification

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
<p>linear equations, iterative methods, error estimates, Poisson equation, over-relaxation</p> <p style="text-align: right;">iB</p>						

**A SURVEY OF ERROR ESTIMATES FOR ITERATIVE SOLUTIONS OF SYSTEMS OF  
LINEAR EQUATIONS WITH AN APPLICATION TO THE SOLUTION OF POISSON'S EQUATION**

Prepared by  
Michael J. Vander Vorst

**ABSTRACT:** In the solution of a set of linear equations  $Ax = b$  by an iterative method one would like an estimate of the error  $x - x_n$  so that the iteration can be stopped after the error is within acceptable bounds. We present a review of several types of error estimates paying particular attention to estimates of the form  $\|x_{n+1} - x\| \leq c \|x_{n+1} - x_n\|$  for the class of iterative methods of the form  $x_{n+1} = Mx_n + k$ . In addition we give a new error estimate for the successive over-relaxation method which is a generalization of an estimate of Sassenfeld for the Gauss-Seidel method. We present a numerical example of an error estimate of Albrecht for the successive over-relaxation method applied to the iterative solution of the system of linear equations which arise from the finite difference formulation of the Poisson equation.

U. S. NAVAL ORDNANCE LABORATORY  
WHITE OAK, MARYLAND

WOLTR 72-189

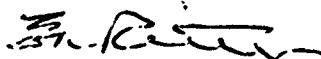
3 October 1972

**A Survey of Error Estimates for Iterative Solutions of Systems of Linear Equations with an Application to the Solution of Poisson's Equation.**

This report is the result of an effort to find acceptable stopping conditions for the iterative solutions of systems of linear equations, including some classes of such systems which arise in the numerical solution of hydrodynamic phenomena. An application is made to the calculation of the pressure about a cavity resulting from an underwater explosion.

This work was sponsored under Task number MAT-03L-000/ZRU0-001-010.

ROBERT WILLIAMSON II  
Captain, USN  
Commander



E. K. RITTER  
By direction

CONTENTS

	Page
INTRODUCTION.....	1
NOTATION AND BASIC CONCEPTS.....	2
ERROR ESTIMATES OF TYPE I.....	5
ERROR ESTIMATES OF TYPE II.....	6
COMMENTS ON THE ERROR ESTIMATES OF TYPE II.....	16
ERROR ESTIMATES OF TYPE III.....	17
NUMERICAL EXAMPLE--SOLUTION OF THE POISSON EQUATION.....	18
CONCLUSION.....	29
ACKNOWLEDGEMENTS.....	33

ILLUSTRATIONS

Figure	Title	Page
1	Cell Nomenclature	20
2	The Computing Mesh	21
3	Example of Particles and Cells	23
4	Cell Flagging for Problems 1 and 2	27
5	Cell Flagging for Problems 3 and 4	28

TABLES

Table	Title	Page
1	Definition of Mesh for Problems 1 and 3	26
2	Parameters Needed to Calculate Error Estimates	30
3	Errors for Problems 1 and 2	31
4	Errors for Problems 3 and 4	32

# INTRODUCTION

Let

$$(1) \quad Ax = b$$

denote a system of  $m$  linear equations in  $m$  unknowns, and let

$$(2) \quad x_{n+1} = Mx_n + k$$

be an iterative method for finding the solution to (1). There are two central questions in using such an iterative method to approximate the solution to a set of linear simultaneous algebraic equations:

- (i) Does the iterative method (2) ultimately converge to the solution of (1), i.e. does

$$\lim_{n \rightarrow \infty} x_n = x ?$$

- (ii) What is the error  $e_n$  or an error estimate  $E_n$  for the difference between the computed solution and the true solution, i.e. what is  $E_n$  such that

$$\| e_n \| = \| x - x_n \| \leq E_n ?$$

The first of these two questions has received considerable exposure in the literature. The texts by Varga <sup>(16)</sup> and Young <sup>(21)</sup> give convergence criteria for a variety of methods; moreover, each of these references has an extensive bibliography on the subject.

Error estimates, on the other hand, have received much less attention. The results that have been derived can be broken down into three categories:

TYPE I. The error estimate  $E_n$  assumes a knowledge of the inverse or an approximate inverse of  $A$ .

TYPE II. The error estimate  $E_n$  is given only in terms of previous iterates, for example suppose there exists a computable constant  $\alpha$  such that for some vector norm

$$\| e_n \| = \| x - x_n \| \leq E_n = \alpha \| x_n - x_{n-1} \|.$$



TYPE III. The error is given in terms of two vectors  $u_n$  and  $v_n$  which bound both the solution  $x$  and the  $n$ -th iterate  $x_n$ , i.e. at each iteration  $u_n$  and  $v_n$  are calculated such that for some partial ordering,  $\leq$ , between vectors

$$u_n \leq x_n \leq v_n \text{ and } u_n \leq x \leq v_n.$$

In this paper we will discuss each of these three types of error bounds. Special emphasis will be placed on the estimates of Type II since the estimates of Type I have recently been thoroughly reviewed by Fitzgerald <sup>(8)</sup>, and the estimates of Type III generally apply only to very specialized problems. We will review the articles that have appeared in the literature which give error estimates of Type II, beginning with the historically interesting paper of von Mises and Pollaczek-Geiringer (17), chronologically proceeding to the latest papers available, and concluding with a numerical example using the successive over-relaxation (SOR) method and an error estimate of Albrecht <sup>(1)</sup>.

#### NOTATION AND BASIC CONCEPTS

Throughout this paper we will consistently use the terminology of Householder <sup>(9)</sup> and Varga <sup>(16)</sup>. For completeness the definitions and concepts used in this paper are presented below.

Let  $R$  denote the real numbers,  $C$  the complex numbers,  $E^m$  the  $m$ -dimensional vector space over the field of complex numbers, and  $G(E^m)$  the set of all  $m \times m$  complex matrices. Then  $\| \cdot \|: E^m \rightarrow R$  is a vector norm on  $E^m$  if

- (a)  $\|x\| \geq 0$  for all  $x$ , and  $\|x\| = 0$  if and only if  $x = 0$
- (b)  $\|\lambda x\| = |\lambda| \cdot \|x\|$ ,  $\lambda$  in  $C$
- (c)  $\|x + y\| \leq \|x\| + \|y\|$ ,  $x, y$  in  $E^m$ .

Furthermore  $\| \cdot \|: G(E^m) \rightarrow R$  is a matrix norm if, in addition to (a), (b), (c),

- (d)  $\|AB\| \leq \|A\| \cdot \|B\|$  for all  $A$  and  $B$  in  $G(E^m)$  is

satisfies. Let  $|| \cdot ||$  be any vector norm on  $E^m$  then the equation

$$|| A || = \sup_{x \neq 0} \frac{|| Ax ||}{|| x ||}$$

defines a matrix norm on  $G(E^m)$  which is said to be induced by the vector norm  $|| \cdot ||$  on  $E^m$ . A matrix norm  $|| \cdot ||$  is consistent with a vector norm  $|| \cdot ||$  if

$$|| Ax || \leq || A || \cdot || x ||$$

for all  $A$  in  $G(E^m)$  and all  $x$  in  $E^m$ ; moreover the matrix norm induced by a vector norm is consistent with that vector norm. Let  $\rho(A)$  denote the spectral radius of  $A$ , then  $\rho(A) \leq || A ||$  for any matrix norm.

Now let  $x^*$  and  $A^*$  respectively denote the conjugate transpose of the vector  $x$  and the matrix  $A$ , where  $A = (a_{ij})$  and  $x = (x_i)$ . Three widely used vector norms and their induced matrix norms are:

$$|| x ||_1 = \sum_{i=1}^m | x_i |, \quad || A ||_1 = \max_{1 \leq j \leq m} \sum_{i=1}^m | a_{ij} |$$

$$|| x ||_2 = \sqrt{x^* x} = \sqrt{\sum_{i=1}^m | x_i |^2}, \quad || A ||_2 = \sqrt{\rho(A^* A)}$$

$$|| x ||_\infty = \max_{1 \leq i \leq m} | x_i |, \quad || A ||_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^m | a_{ij} |$$

It is well known (Varga <sup>(16)</sup>) that the sequence  $x_n$  from (2) converges for any  $x_0$  if and only if  $\rho(M) < 1$ . If  $A$  is nonsingular and the method (2) is derived from the system (1), i.e.  $x = A^{-1}b$  is the only solution to  $x = Mx + k$ , then  $x = \lim_{n \rightarrow \infty} x_n$

is the unique solution to (1).

The three most commonly used iterative methods are the Jacobi\*, Gauss-Seidel\*\*, and successive over-relaxation methods. Each of these three methods will now be expressed in the form (2). Decompose A into a diagonal matrix D, a lower triangular matrix E, and an upper triangular matrix F such that

$$(3) \quad A = D - E - F,$$

then the Jacobi method

$$(4) \quad x_{i,(n+1)} = \frac{1}{a_{ii}} \left[ - \sum_{\substack{j=1 \\ j \neq i}}^m a_{ij} x_{j,(n)} + b_i \right]$$

can be written

$$x_{n+1} = D^{-1}(E + F)x_n + D^{-1}b$$

so that the Jacobi matrix B is  $D^{-1}(E + F)$ . The Gauss-Seidel method

$$(5) \quad x_{i,(n+1)} = \frac{1}{a_{ii}} \left[ - \sum_{j=1}^{i-1} a_{ij} x_{j,(n+1)} - \sum_{j=i+1}^m a_{ij} x_{j,(n)} + b_i \right]$$

can be written

$$x_{n+1} = (D - E)^{-1}Fx_n + (D - E)^{-1}b$$

so that the Gauss-Seidel matrix  $L_1$  is  $(D - E)^{-1}F$ . The successive over-relaxation method

$$(6) \quad x_{i,(n+1)} = \frac{\omega}{a_{ii}} \left[ - \sum_{j=1}^{i-1} a_{ij} x_{j,(n+1)} - \sum_{j=i+1}^m a_{ij} x_{j,(n)} + b_i \right] - (\omega-1)x_{i,(n)}$$

can be written

$$x_{n+1} = (D - \omega E)^{-1} [(1 - \omega)D + \omega F] x_n + \omega(D - \omega E)^{-1}b$$

---

\*Jacobi method is also called the point total step method, the method of simultaneous displacements, and the Richardson iterative method. Note the total step method is a translation of the German word Gesamtschrittverfahren.

\*\*The Gauss-Seidel method is also called the point single step method, the method of successive displacements, and the Liebmann method. Note that single step method is a translation of the German word Einzelshrittverfahren.

so that the successive over-relaxation matrix  $L_\omega$  is  $(D - \omega E)^{-1} [(1 - \omega)D + \omega F]$ . We note that for  $\omega = 1$  the successive over-relaxation method reduces to the Gauss-Seidel method. Young <sup>(20)</sup> shows that under certain conditions there exists an optimum over-relaxation parameter,  $1 \leq \omega \leq 2$ , such that  $\rho(L_\omega) < 1$ , and that the successive over-relaxation method then converges faster than either the Jacobi or Gauss-Seidel methods.

We now proceed to discuss the error estimates of types I, II, and III.

#### ERROR ESTIMATES OF TYPE I

Recently Fitzgerald <sup>(8)</sup> developed bounds for the error in a computed inverse of a matrix and for the approximate solution of  $Ax = b$ . These bounds are of the first type listed above in that he assumes the availability of an approximate inverse  $X$  of  $A$ . Methods in which these bounds would be most applicable compute an approximate inverse  $X$  of  $A$  and let  $y = Xb$ . The residual  $I - XA$  or  $I - AX$  is then computed and from these a bound on the error  $e = x - y$  is found using the well known inequality

$$\|y - x\| \leq \|A^{-1}\| \cdot \|Ay - b\|$$

where

$$\|A^{-1}\| \leq \frac{\|X\|}{1 - \|I - AX\|}$$

if  $\|I - AX\| < 1$ .

If this error is too large  $X$  may be improved until the error is within acceptable bounds. When one uses such a method to solve for the solution vector  $x$  of  $AX = b$  for just one  $b$ , he is actually computing much more than he needs, namely  $A^{-1}$ .

Fitzgerald feels that in general it is futile to expect to find a good error bound without some knowledge of the inverse of  $A$ ; however, one would hope that if  $A$  had a "special enough" structure it would be possible to find an error estimate without knowing anything about  $A^{-1}$ , that is find error bounds of the second

or third types above. As we shall see, this is true for some frequently used methods employed on classes of matrices which often arise in the solution of problems derived from physical phenomena.

#### ERROR ESTIMATES OF TYPE II

We now present the error estimates of the form

$$|| x - x_n || \leq \alpha || x_n - x_{n-1} ||.$$

Although some of these bounds were originally neither presented nor proved in this form, for simplicity and self-consistency they will be given using the results of the following theorem due to Weissinger<sup>(19)</sup>.

Theorem I. Let  $|| \cdot ||$  be a matrix norm consistent with the vector norm  $|| \cdot ||$ , and let  $x_{n+1} = Mx_n + k$  be a method derived from the nonsingular system  $Ax = b$ . If  $|| M || < 1$  then

$$x = \lim_{n \rightarrow \infty} x_n, \text{ and}$$

$$(7) \quad || x - x_n || \leq \frac{|| M ||}{1 - || M ||} || x_n - x_{n-1} ||$$

Proof: Since  $\rho(M) \leq || M || < 1$  the method converges to the unique solution of  $AX = b$ . For the proof of the second statement let

$$e_n = x - x_n, \text{ and}$$

$$\delta_n = x_n - x_{n-1}.$$

Then since  $\delta_{n+1} = M\delta_n$  and  $\rho(M) < 1$ , it can be easily shown that

$$e_n = (I - M)^{-1} M \delta_n,$$

and since  $|| \cdot ||$  is a consistent matrix norm

$$|| x - x_n || \leq || (I - M)^{-1} || \cdot || M || \cdot || x_n - x_{n-1} ||.$$

Now again since  $\rho(M) < 1$  we can expand  $(I - M)^{-1}$  as  $\sum_{i=0}^{\infty} M^i$ , then using

$$\| (I - M)^{-1} \| \leq \sum_{i=0}^{\infty} \| M \|^i = \frac{1}{1 - \| M \|}$$

we have the result of the theorem.

We can see from Theorem I that if we can find an easily calculable, consistent matrix norm such that  $\| M \| < 1$ , then we have a computable error bound. This is in fact what has been done for the Jacobi, Gauss-Seidel, and successive over-relaxation methods using the three norms  $\| \cdot \|_1$ ,  $\| \cdot \|_2$ , and  $\| \cdot \|_{\infty}$ . We must note however that in general any or all of the above norms may give  $\| M \| \geq 1$ , precluding the use of the error bound (7), but the method may still converge.

We now present the error bounds of Type II that have been derived. In all that follows we assume we are solving the m-th order system  $Ax = b$ , where  $A = (a_{ij})$ ,  $b = (b_i)$ , and  $x = (x_i)$ , by the method  $x_{n+1} = Mx_n + k$  where  $x_n = (x_{i,(n)})$  and  $M$  is matrix of order m which may be one of the matrices  $B$ ,  $L_1$ , or  $L_{\omega}$  corresponding to the Jacobi, Gauss-Seidel, or successive over-relaxation methods respectively.

In 1929 von Mises and Pollaczek-Geiringer<sup>(17)</sup> proved that if

$$\sum_{\substack{i=1 \\ i \neq j}}^m \left| \frac{a_{ij}}{a_{ii}} \right| \leq \mu < 1 \text{ for } j = 1, \dots, m$$

then an error estimate for the Jacobi method is

$$\sum_{i=1}^m \left| x_{i,(n+1)} - x_i \right| \leq \frac{\mu}{1-\mu} \sum_{i=1}^m \left| x_{i,(n+1)} - x_{i,(n)} \right|.$$

Equivalently using the terminology and the result of Theorem I we have:

If  $\mu = \|B\|_1 < 1$ , then

$$\|x_{n+1} - x\|_1 < \frac{\mu}{1-\mu} \|x_{n+1} - x\|_1.$$

No other results were obtained until 1942 when Collatz<sup>(5)</sup> showed that for the Jacobi method, if

$$\sum_{j=1, j \neq i}^m \left| \frac{a_{ij}}{a_{ii}} \right| \leq \xi < 1, \text{ for } i=1, \dots, m$$

then

$$\max_{1 \leq i \leq m} |x_{i,(n+1)} - x_i| \leq \frac{\xi}{1-\xi} \max_{1 \leq i \leq m} |x_{i,(n+1)} - x_{i,(n)}|$$

He also noted the same bound held for the Gauss-Seidel method. Again we restate this bound in the form of Theorem I to get: If  $\xi = \|B\|_\infty < 1$  then

$$\|x_{n+1} - x\|_\infty \leq \frac{\xi}{1-\xi} \|x_{n+1} - x_n\|_\infty.$$

Again there was a long lapse until 1951 when Sassenfeld<sup>(11)</sup> presented two results for the Gauss-Seidel method, namely,

Criterion I. Let

$$\alpha_i = \frac{1}{a_{ii}} \sum_{j=1}^{i-1} |a_{ij}| \alpha_j + \sum_{j=i+1}^m |a_{ij}|, \quad i=1, \dots, m$$

and  $\alpha = \max_i \alpha_i$ . If  $\alpha < 1$  then

$$\max_i |x_{i,(n+1)} - x_i| \leq \frac{\alpha}{1-\alpha} \max_i |x_{i,(n+1)} - x_{i,(n)}|.$$

Criterion II. Let

$$\beta_i = \frac{1}{|a_{ii}|} \left[ \sum_{j=1}^{i-1} |a_{ij}| \left( \max_{k < i} \beta_k \right) + \sum_{j=i+1}^m |a_{ij}| \right], \quad i=1, \dots, m$$

and  $\beta = \max_i \beta_j$ . If  $\beta < 1$ , then

$$\max_i \left| x_{i,(n+1)} - x_i \right| \leq \frac{\beta}{1-\beta} \max_i \left| x_{i,(n+1)} - x_{i,(n)} \right|$$

We observe that Criterion I is the stronger of the two since  $\alpha \leq \beta$ . Both of Sassenfeld's criteria will be proven as corollaries to the following more general theorem on the successive over-relaxation method which uses Theorem I with an explicit representation of  $L_\omega$  and the norm  $\|\cdot\|_\infty$ . Referring to (6) we can write the successive over-relaxation matrix  $J_\omega = (s_{ij})$  as

$$s_{11} = -(\omega - 1)$$

$$s_{1j} = \frac{\omega}{a_{11}} a_{1j}, j=2, \dots, m$$

and for  $i=2, \dots, m$

$$s_{ij} = -\frac{\omega}{a_{ii}} \sum_{k=1}^{i-1} a_{ik} s_{kj}, j=1, \dots, i-1$$

$$s_{ii} = -\frac{1}{a_{ii}} \left[ (\omega - 1)a_{ii} + \omega \sum_{k=1}^{i-1} a_{ik} s_{kj} \right]$$

$$s_{ij} = -\frac{\omega}{a_{ii}} (a_{ij} + \omega \sum_{k=1}^{i-1} a_{ik} s_{kj}), j=i+1, \dots, m.$$

Theorem II. Let  $\gamma_i = \sum_{j=1}^m |s_{ij}|$  and  $\gamma = \max_i \gamma_i$ , then if  $\gamma < 1$ , an error estimate for the successive over-relaxation method is

$$\max_i \left| x_{i,(n+1)} - x_i \right| \leq \frac{\gamma}{1-\gamma} \max_i \left| x_{i,(n+1)} - x_{i,(n)} \right|$$

Proof: We note that  $\gamma = \|L_\omega\|_\infty$ . Then if  $\gamma < 1$  we have

$$\left\| x_{n+1} - x \right\|_\infty \leq \frac{\gamma}{1-\gamma} \left\| x_{n+1} - x_n \right\|_\infty$$

which is the conclusion of the theorem.



**Corollary I.** Let

$$\delta_1 = |\omega - 1| + \frac{\omega}{a_{11}} \sum_{j=1}^m |a_{1j}|$$

$$\delta_i = \frac{1}{|a_{ii}|} \left[ \omega \sum_{j=1}^{i-1} |a_{ij}| \delta_j + |(\omega - 1)a_{ii}| \right. \\ \left. + \omega \sum_{j=i+1}^m |a_{ij}| \right], i=2, \dots, m$$

and  $\delta = \max_i \delta_i$ .

If  $\delta < 1$  then for the successive over-relaxation method, we have

$$\max_i |x_{i,(n+1)} - x_i| \leq \frac{\delta}{1-\delta} \max_i |x_{i,(n+1)} - x_{i,(n)}|.$$

**Proof:** We need only to show that  $\gamma \leq \delta$ . For  $i > 1$  we can write

$$\gamma_i = \sum_{j=1}^m |s_{ij}|$$

$$= \frac{1}{|a_{ii}|} \left[ \omega \sum_{j=1}^{i-1} \left| \sum_{k=1}^{i-1} a_{ik} s_{kj} \right| + \left| (\omega - 1)a_{ii} + \omega \sum_{k=1}^{i-1} a_{ik} s_{kj} \right| \right. \\ \left. + \omega \sum_{j=i+1}^m \left| a_{ij} + \omega \sum_{k=1}^{i-1} a_{ik} s_{kj} \right| \right]$$

$$\leq \frac{1}{|a_{ii}|} \left[ \omega \sum_{k=1}^{i-1} |a_{ik}| \sum_{j=1}^m |s_{kj}| + |(\omega - 1)a_{ii}| + \omega \sum_{j=i+1}^m |a_{ij}| \right]$$

$$= \frac{1}{|a_{ii}|} \left[ \omega \sum_{j=1}^{i-1} |a_{ij}| \gamma_j + |(\omega - 1)a_{ii}| + \omega \sum_{j=i+1}^m |a_{ij}| \right]$$

Now  $\delta_1 = \gamma_1$ , hence  $\gamma_i \leq \delta_i$  and  $\gamma \leq \delta$ .

Corollary II. Let

$$v_1 = \delta_1$$

$$v_i = \frac{1}{|a_{ii}|} \left[ \omega \left( \max_{j < i} \delta_j \right) \sum_{j=1}^{i-1} |a_{ij}| + |(\omega - 1) a_{ii}| \right. \\ \left. + \omega \sum_{j=i+1}^n |a_{ij}| \right], i=2, \dots, n$$

and  $v = \max_i v_i$ .

If  $v < 1$  then for the successive over-relaxation method, we have

$$\max_i |x_{i,(n+1)} - x_i| \leq \frac{v}{1-v} \max_i |x_{i,(n+1)} - x_{i,(n)}|.$$

Proof: This corollary obviously follows from Corollary I since  $v < \delta$ .

Corollary III. Criterion I and II of Sassenfeld.

Proof: Let  $\omega = 1$  in Corollaries I and II.

Dueck <sup>(6)</sup> presented the following error estimate for the Gauss-Seidel method which is slightly better than Collatz's estimate but not as good as Sassenfeld's.

Let  $A_2$  be the upper triangular part of the Jacobi matrix  $B$ . If  $\|B\|_\infty < 1$  then

$$\|x_{n+1} - x\|_\infty \leq \frac{\|A_2\|_\infty}{1 - \|B\|_\infty} \|x_{n+1} - x_n\|_\infty.$$

Let  $A_1$  be the lower triangular part of  $B$ , then Dueck proved this estimate by noting that an equivalent formulation of the Gauss-Seidel method is

$$x_{n+1} - A_1 x_{n+1} = A_2 x_n + k$$

and

$$x - x_{n+1} = A_1(x - x_{n+1}) + A_2(x - x_n)$$

$$(I - A_1)(x - x_{n+1}) = A_2(x - x_n)$$

$$= A_2(x - x_{n+1}) + A_2(x_{n+1} - x_n)$$

$$(I - A_1 - A_2)(x - x_{n+1}) = A_2(x_{n+1} - x_n)$$

or 
$$x - x_{n+1} = (I - B)^{-1} A_2(x_{n+1} - x_n).$$

Therefore by Theorem I

$$\|x - x_{n+1}\|_{\infty} \leq \frac{\|A_2\|_{\infty}}{1 - \|B\|_{\infty}} \|x_{n+1} - x_n\|_{\infty}.$$

Feldman<sup>(7)</sup> found an error estimate for the Gauss-Seidel method which is comparable to that of Dueck. As before let  $A_1$  be the lower triangular part and let  $A_2$  be the upper triangular part of  $B$ . Let

$$\psi = \begin{cases} \frac{1 - \|A_1\|_{\infty}^m}{1 - \|A_1\|_{\infty}} \|A_2\|_{\infty} & \text{if } \|A_1\|_{\infty} \neq 1 \\ m\|A_2\|_{\infty} & \text{if } \|A_1\|_{\infty} = 1. \end{cases}$$

If  $\psi < 1$  an error estimate for the Gauss-Seidel method is

$$\|x_{n+1} - x_n\|_{\infty} \leq \frac{\psi}{1-\psi} \|x_{n+1} - x_n\|_{\infty}.$$

Following the proof given by Feldman, we have

$$L_1 = (I - A_1)^{-1} A_2$$

and

$$\|L_1\|_{\infty} \leq \|(I - A_1)^{-1}\|_{\infty} \|A_2\|_{\infty}.$$

However since  $I - A_1$  is lower triangular we can write

$$(I - A_1)^{-1} = I + A_1 + A_1^2 + \dots + A_1^{n-1}$$

and

$$\| (I - A_1)^{-1} \|_{\infty} \leq \sum_{i=0}^{n-1} \| A_1 \|_{\infty}^i$$

and

$$\| L_1 \|_{\infty} \leq \left( \sum_{j=0}^{n-1} \| A_1 \|_{\infty}^j \right) \| A_2 \|_{\infty}.$$

Then using Theorem I we have  $\geq$  Feldman's result.

Albrecht<sup>(1)</sup> derived an error estimate for the successive over-relaxation method for the important case when  $A$  is Hermitian, positive definite, and 2-cyclic. [ $A$  is 2-cyclic if there is a permutation of its Jacobi matrix  $B$  such that

$$PBP^T = \begin{pmatrix} 0 & A_{12} \\ A_{21} & 0 \end{pmatrix}$$

where the zero blocks are square.] We first transform the system (1) into a similar system whose Jacobi matrix is Hermitian. As in (3) let  $A = D - E - F$  where  $D$  is diagonal,  $E$  is lower triangular, and  $F$  is upper triangular. Let

$$T = D^{-\frac{1}{2}}(E + F)D^{-\frac{1}{2}},$$

$$a = D^{-\frac{1}{2}}b, \text{ and}$$

$$y = D^{\frac{1}{2}}x,$$

then

$$y_{n+1} = Ty_n + a$$

is the Jacobi method for the solution of

$$(8) \quad (I - T)y = a$$

and the solution of  $Ax = b$  is  $x = D^{-\frac{1}{2}}y$ . However  $T$  like  $A$  is Hermitian and positive definite while the Jacobi matrix  $B$  is generally not. Using the above notation, Albrecht's error estimate for the solution to (8) by the successive

over-relaxation method when A is Hermitian, positive definite, and 2-cyclic is

$$\|y_{n+1} - y\|_2 \leq \lambda \|y_{n+1} - y_n\|_2$$

where

$$\lambda = \frac{\frac{1}{2} \sqrt{\|T\|_2^2 (\rho^2 + \|T\|_2^2) + 4\gamma^2 (1 - \|T\|_2^2)} + \sqrt{\|T\|_2^2 (\rho^2 + \|T\|_2^2)}}{1 - \|T\|_2^2}$$

and

$$\gamma = 1 - \frac{1}{\omega} \text{ and } \rho = \frac{2}{\omega} - 1.$$

Now since  $x = D^{-\frac{1}{2}}y$ ,  $D = \text{diagonal}(d_i)$ , an error estimate for the solution of  $Ax = b$  is

$$(9) \quad \|x_{n+1} - x\|_2 \leq \sqrt{4 \max_i \frac{1}{d_i}} \|y_{n+1} - y\|_2$$

The only thing remaining is to calculate  $\|T\|_2$ , but  $T$  is Hermitian hence  $\|T\|_2 = \rho(T)$ . Varga<sup>(16, section 9)</sup> gives methods for finding  $\rho(T)$  which are based on the iterative scheme (2) and are thus quite easy to implement since the same scheme is used to solve the set of equations.

Young<sup>(20)</sup> shows that in the case of 2-cyclic matrices (i.e. matrices with Young's Property A) if the spectral radius  $\rho$  of the Jacobi matrix is less than one then the optimum over-relaxation factor  $\omega$  is related to  $\rho$  by

$$\omega = \frac{2}{1 + \sqrt{1 - \rho^2}}$$

and

$$\rho(L) = \omega - 1,$$

where  $L$  is the associated over-relaxation matrix. Hence if  $\rho < 1$  we have  $0 \leq \omega < 2$  and  $\rho(L) < 1$ . Now in our case where  $A$  is Hermitian, positive definite, and 2-cyclic, the over-relaxation matrix  $P_\omega$  for the system (8) is similar to the matrix  $L_\omega$  (see (6)) for the system (1) by

$$P_\omega = D^{-\frac{1}{2}} L_\omega D^{-\frac{1}{2}}$$

and

$$T = D^{\frac{1}{2}}BD^{-\frac{1}{2}}$$

therefore  $\rho(P_u) = \rho(L_u)$  and  $\rho(T) = \rho(B)$ .

Similarly denote the constant vector  $k$  of (2) by  $k_1$  for  $L_u$  and by  $k_p$  for  $P_u$ , then

$k_p = D^{\frac{1}{2}}k_1$ . If as before we let

$$y_0 = D^{\frac{1}{2}}x_0$$

$$x_{n+1} = L_u x_n + k_1,$$

and

$$y_{n+1} = P_u y_n + k_p,$$

then

$$y_n = D^{\frac{1}{2}}x_n.$$

Therefore by (9) an error estimate for the method  $x_{n+1} = L_u x_n + k_1$  is

$$(10) \quad ||x_{n+1} - x||_2 \leq \lambda \sqrt{\max_i \frac{1}{d_i}} ||D^{\frac{1}{2}}(x_{n+1} - x_n)||_2,$$

or a less desirable estimate is

$$(11) \quad ||x_{n+1} - x||_2 \leq \lambda \sqrt{(\max_i \frac{1}{d_i})(\max_i d_i)} ||x_{n+1} - x_n||_2.$$

To use Albrecht's estimate we have two alternatives. We can either solve the system (8) for  $y$  using the error estimate (9), and after sufficient convergence let  $x = D^{-\frac{1}{2}}y$ , or we can solve the original system (1), using the error estimate (1). The two methods are equivalent and neglecting any computational aspects such as round-off error should give the same results. In the example at the end of this paper we decided to take the second approach only for the reason that for our case it was easier to program.

We now turn to a rather singular result by Weinberger<sup>(19)</sup> which gives an error bound whose range is determined from a maximum principle. Specifically if the real matrix  $M$  of the iterative method  $x_{n+1} = Mx_n + k$  is symmetric let

$$\delta_n = x_n - x_{n-1},$$

$$e_n = x - x_n,$$

$$\alpha = || \delta_{n-1} ||_2^2,$$

$$\beta = \delta_{n-1}^T \delta_n,$$

and

$$\gamma = || \delta_n ||_2^2,$$

then if  $\rho(M) \leq 1 - \epsilon < 1$  the range of possible values of  $|| e_n ||_2^2$

is equal to the range of

$$R(\psi) = \frac{(\alpha\gamma - \beta^2)\psi^4}{(1 - \psi)^2} + \frac{(\beta\psi - \gamma)^4}{[(\alpha - \beta)\psi - (\beta - \gamma)]^2} \bigg/ [\alpha\psi^2 - 2\beta\psi + \gamma]$$

on the interval

$$\frac{(1 - \epsilon)\beta + \gamma}{(1 - \epsilon)\alpha + \beta} \leq \psi \leq 1 - \epsilon.$$

The requirement that  $M$  be symmetric generally limits this result to the solution of real linear systems by the Jacobi method

$$y_{n+1} = Ty_n + a$$

derived from the transformed system (8).

Finally Schroeder<sup>(13)</sup> used a theorem concerning an abstract iteration process  $u_{n+1} = Zu_n$  to derive error bounds for the Jacobi and Gauss-Seidel methods which are essentially the same as the bounds of Collatz and Sassenfeld respectively.

#### COMMENTS ON THE ERROR ESTIMATES OF TYPE II

All of the estimates except Albrecht's and possibly Weinberger's only hold if  $|| M || < 1$  where  $|| \cdot ||$  is some consistent matrix norm and  $M$  is the matrix from the iterative process  $x_{n+1} = Mx_n + k$ . However the condition for convergence is  $\rho(M) < 1$ , whereas it is possible to have  $\rho(M) < 1 \leq || M ||$ . We would then have a convergent method but no usable error estimate. In particular for the

bounds of von Mises and Pollaczek-Geiringer, Collatz, Sassenfeld, the author's Theorem II, and Dueck, it is necessary that the matrix  $A$  of  $Ax = b$  be strictly diagonally dominant, i.e.

$$\sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| < |a_{ii}|, \quad i=1, \dots, n.$$

This is unfortunate since the matrices that are generated by many physical problems, e.g. the numerical solution of elliptic differential equations, do not enjoy this property.

On the other hand Albrecht's estimate for the successive over-relaxation method relies only on the spectral radius of the matrix  $M$ , but it requires that  $A$  be positive definite, Hermitian, and 2-cyclic. In the example at the end of this paper we solve a set of linear equations derived from the numerical solution of Poisson's equation by the successive over-relaxation method. For this example the only estimate that is applicable is Albrecht's.

#### ERROR ESTIMATES OF TYPE III

Schroeder<sup>(15)</sup> and Albrecht<sup>(2),(3),(4)</sup> give error estimates of Type III for monotone iterative methods. A description of either monotone methods themselves or conditions under which such a method will produce a monotone sequence of vectors which converge to the solution of the system of linear equations is beyond the scope of this paper. However, Schroeder<sup>(14)</sup> gives a derivation of monotone methods as well as sufficient conditions for the convergence of these methods.

Given a monotone iterative method, it is usually easy to write error bounds for this method. For example let  $\leq$  be the component-wise partial ordering between vectors,  $x$  the solution vector of the set of equations, and  $x_0$  and  $y_0$  be two vectors such that  $x_0 \leq x \leq y_0$ , then if the method gives successive iterates



$x_1$  and  $y_1$  such that

$$x_c \leq x_1 \leq \dots \leq x_n \leq x \leq y_n \leq \dots \leq y_1 \leq y_0$$

then an error estimate for  $x - x_n$  or  $y - y_n$  is

$$|x_i - x_{i,(n)}| \leq |y_{i,(n)} - x_{i,(n)}|, \text{ for } i=1, \dots, m.$$

This example is not meant to give the most general estimates for all monotone methods. It is however representative of the type of error estimates obtainable with these methods.

There are a few distinct disadvantages to monotone iterative methods. First, the range of applicability of these methods is small; second, it is usually difficult to find initial values with the desired properties; and third, the iteration itself is more complicated, often as in the case of our example requiring two or more sequences which converge to the solution. On the other hand, termination criteria for such iterations are easily determined as the iterates give both upper and lower bounds on the solution.

#### NUMERICAL EXAMPLE--SOLUTION OF THE POISSON EQUATION

In this section we present an example of Albrecht's error estimate for the solution of the set of linear equations derived from the discretized Poisson's equation. The error estimate was programmed into the computer code MACNOL (Marker and Cell Method of the Naval Ordnance Laboratory) which solves incompressible, viscous, initial value, fluid flow problems by the marker and cell method. The MACNOL code is a modification of the MACYL code of Pritchett<sup>(10)</sup>. One modification of MACYL which was incorporated into MACNOL was to reprogram the routine which solves the discretized Poisson's equation to use the successive over-relaxation method instead of the Gauss-Seidel method.

As a note of interest we would like to report that for the sample problem we solved, the successive over-relaxation method converged about 100 times faster than the Gauss-Seidel method.

A brief derivation of the set of linear equations arising from the finite difference solution of the Poisson equation will now be given. Using a cylindrical coordinate system let  $r$  be the radial dimension and  $z$  be the vertical dimension. We denote a finite difference cell by the indices  $i$  and  $j$  where  $i$  varies with radius and  $j$  varies with height.

Figure 1 illustrates this nomenclature:

$\Delta r_i$  = radial dimension of cell  $i,j$

$\Delta z_j$  = vertical dimension of cell  $i,j$

$r_i$  = distance from axis to center of cell  $i,j$

$z_j$  = distance from bottom of mesh to center of cell  $i,j$

$r_{i-\frac{1}{2}}$  = distance from axis to inner boundary of cell  $i,j$

$r_{i+\frac{1}{2}}$  = distance from axis to outer boundary of cell  $i,j$

$z_{j-\frac{1}{2}}$  = distance from bottom of mesh to lower boundary of cell  $i,j$

$z_{j+\frac{1}{2}}$  = distance from bottom of mesh to upper boundary of cell  $i,j$

$\Delta r_{i-\frac{1}{2}} = r_i - r_{i-1}$

$\Delta r_{i+\frac{1}{2}} = r_{i+1} - r_i$

$\Delta z_{j-\frac{1}{2}} = z_j - z_{j-1}$

$\Delta z_{j+\frac{1}{2}} = z_{j+1} - z_j$

Figure 2 illustrates the computing mesh. Rows 1 and N and columns 1 and M are physically fictitious. They are used for convenience in representing the

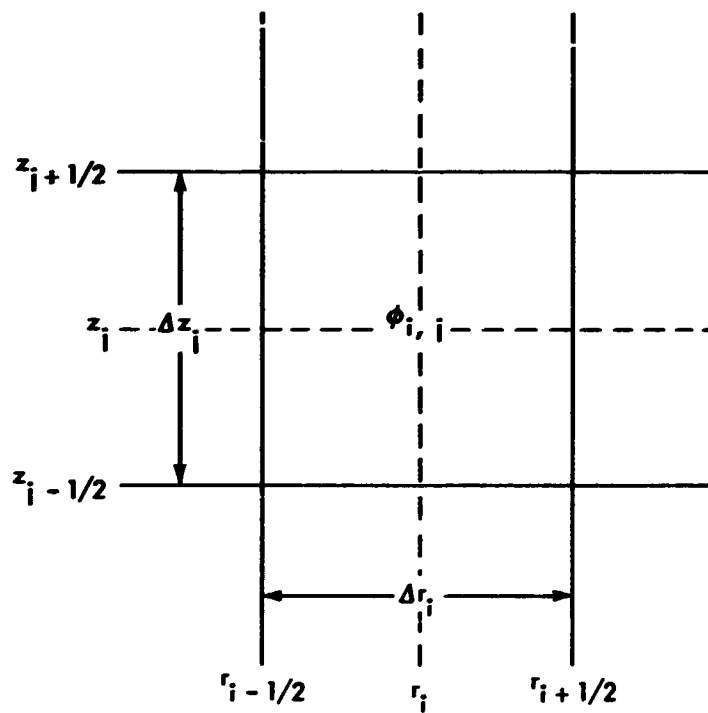


FIG. 1 NOMENCLATURE FOR CELL  $i, j$

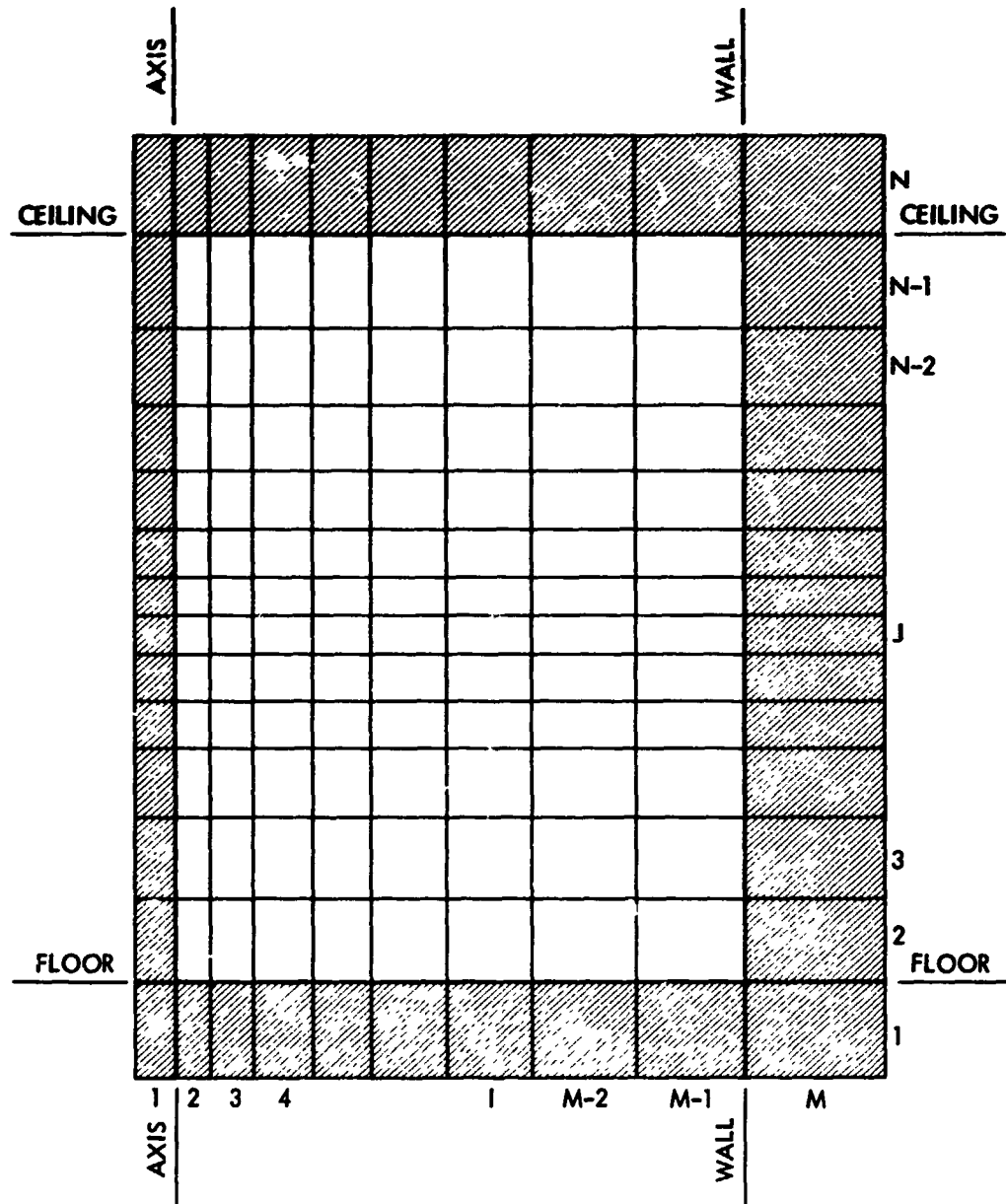


FIG. 2 THE COMPUTING MESH

governing finite difference equations. The domain of the problem need not be defined on each cell of the mesh; by suitable flagging we can represent free surfaces. Figure 3 gives an example of such cell flagging. The computation is only done in full (F) cells; the boundary conditions are applied on the surface (S) cells as well as the axis, wall, floor, and ceiling. The empty (E) cells are only used in the calculation to specify boundary conditions. Let  $\phi$  denote the ratio of pressure to constant density and let  $g$  be the acceleration due to gravity. We wish to solve the Poisson equation in cylindrical coordinates

$$\frac{\partial^2 \phi}{\partial r^2} + \frac{1}{r} \frac{\partial \phi}{\partial r} + \frac{\partial^2 \phi}{\partial z^2} = -R(r, z)$$

with the mixed Neumann-Dirichlet boundary conditions

$$\frac{\partial \phi}{\partial r} = 0 \quad \text{on the axis}$$

$$\frac{\partial \phi}{\partial r} = w(z) \quad \text{on the wall}$$

$$\frac{\partial \phi}{\partial z} = g \quad \text{on the floor and ceiling}$$

and  $\phi$  is prescribed on the surface,

where  $R(r, z)$  and  $w(z)$  are known. Let  $\phi$  be defined at cell centers. The MACNOL code solves the following discretized form of the above problem for  $\phi$

$$(12) \quad \left[ \frac{1}{r_i \Delta r_i} \frac{r_{i+\frac{1}{2}}}{\Delta r_{i+\frac{1}{2}}} (\phi_{i,j} - \phi_{i+1,j}) + \frac{r_{i-\frac{1}{2}}}{\Delta r_{i-\frac{1}{2}}} (\phi_{i,j} - \phi_{i-1,j}) \right] \\ + \frac{1}{\Delta z_j} \left[ \frac{1}{\Delta z_{j+\frac{1}{2}}} (\phi_{i,j} - \phi_{i,j+1}) + \frac{1}{\Delta z_{j-\frac{1}{2}}} (\phi_{i,j} - \phi_{i,j-1}) \right] = -R_{i,j}$$

23

with the boundary conditions

$$\begin{aligned} \phi_{1,j} &= \phi_{2,j}, \\ \phi_{M,j} &= \phi_{M-1,j} + \Delta x_{M-1/2} w_j, \\ (13) \quad \phi_{1,1} &= \phi_{1,2} - g \Delta z_{3/2}, \\ \phi_{1,N} &= \phi_{1,N-1} + g \Delta z_{N-1/2} \end{aligned}$$

and  $\phi_{1,j}$  is given if  $1,j$  is a surface cell, and where  $R_{1,j}$  and  $w_j$  are regarded as known.

Where applicable substitute the boundary conditions (13) into the equations (12). The result is a set of linear equations in the unknowns  $\phi_{i,j}$ , where  $i,j$  ranges over the set of full cells. Denote this set of equations by DPE (Discretized Poisson's Equation). If the set of equations DPE is naturally ordered, i.e. ordered by the rows or columns of the finite difference mesh, then its matrix A is symmetric, positive definite, and 2-cyclic. The matrix A is not, however, strictly diagonally dominant, although it is irreducible and

$$\sum_{\substack{j=1 \\ j \neq i}}^m |a_{ij}| \leq |a_{ii}|, \quad i=1, \dots, m$$

with inequality for at least one  $i$ . That is, A is irreducibly diagonally dominant.

To check the usefulness of Albrecht's estimates for use in the MACNOL code, we programmed four similar sample problems whose solutions are known exactly. The four problems are broken down into two sets of two problems each. The first problem in each set uses an unevenly spaced mesh, the second uses an evenly spaced mesh. The first set of problems is simply

water standing still in a cylindrical tank. For the second set of problems, in order to test the error estimates on a non-rectangular mesh, we introduced the artificial problem of an axi-symmetric underwater cavity whose surface was at hydrostatic pressure. The solution to all of these problems is simply hydrostatic pressure.

To reiterate, Problem 1 uses an unevenly spaced rectangular mesh; Problem 2 uses an evenly spaced rectangular mesh; Problem 3 uses an unevenly spaced non-rectangular mesh; and Problem 4 uses an evenly spaced non-rectangular mesh.

All four problems use a mesh of dimension 56 in the radial direction and 122 in the vertical direction. When the mesh is evenly spaced

$$\Delta r = \Delta z = .5.$$

For the unevenly spaced problems  $\Delta r_1$  and  $\Delta z_1$  are given in Table 1. Figure 4 shows the cell flagging for Problems 1 and 2, and Figure 5 shows the cell flagging for Problems 3 and 4.

The error estimates given are (10) and (11) of Albrecht. Let  $D$  be the diagonal of DPE then from (10) we have

$$(14) \quad || \phi_{n+1} - \phi ||_2 \leq \alpha || D^{1/2} (\phi_{n+1} - \phi_n) ||_2$$

$$\text{where} \quad \alpha = \lambda || D^{-1/2} ||_2$$

and from (11)

$$(15) \quad || \phi_{n+1} - \phi ||_2 \leq \beta || \phi_{n+1} - \phi_n ||_2$$

$$\text{where} \quad \beta = \lambda || D^{-1/2} ||_2 || D^{1/2} ||_2$$

We also investigated the possibility of using

$$(16) \quad || \phi_{n+1} - \phi ||_2 \approx \frac{\rho(L_\omega)}{1 - \rho(L_\omega)} || \phi_{n+1} - \phi_n ||_2 = \delta || \phi_{n+1} - \phi_n ||_2$$



TABLE I DEFINITION OF MESH FOR PROBLEMS 1 AND 3

$R_i + 1/2$	$\Delta R_i$	$i$	$R_i + 1/2$	$\Delta R_i$	$i$	$Z_i + 1/2$	$\Delta Z_i$	$i$
0.0	0.0	1	0.0	0.0	1	0.0	0.0	1
0.1	0.1	2	0.1	0.1	2	0.1	0.1	2
0.2	0.2	3	0.2	0.2	3	0.2	0.2	3
0.3	0.3	4	0.3	0.3	4	0.3	0.3	4
0.4	0.4	5	0.4	0.4	5	0.4	0.4	5
0.5	0.5	6	0.5	0.5	6	0.5	0.5	6
0.6	0.6	7	0.6	0.6	7	0.6	0.6	7
0.7	0.7	8	0.7	0.7	8	0.7	0.7	8
0.8	0.8	9	0.8	0.8	9	0.8	0.8	9
0.9	0.9	10	0.9	0.9	10	0.9	0.9	10
1.0	1.0	11	1.0	1.0	11	1.0	1.0	11
1.1	1.1	12	1.1	1.1	12	1.1	1.1	12
1.2	1.2	13	1.2	1.2	13	1.2	1.2	13
1.3	1.3	14	1.3	1.3	14	1.3	1.3	14
1.4	1.4	15	1.4	1.4	15	1.4	1.4	15
1.5	1.5	16	1.5	1.5	16	1.5	1.5	16
1.6	1.6	17	1.6	1.6	17	1.6	1.6	17
1.7	1.7	18	1.7	1.7	18	1.7	1.7	18
1.8	1.8	19	1.8	1.8	19	1.8	1.8	19
1.9	1.9	20	1.9	1.9	20	1.9	1.9	20
2.0	2.0	21	2.0	2.0	21	2.0	2.0	21
2.1	2.1	22	2.1	2.1	22	2.1	2.1	22
2.2	2.2	23	2.2	2.2	23	2.2	2.2	23
2.3	2.3	24	2.3	2.3	24	2.3	2.3	24
2.4	2.4	25	2.4	2.4	25	2.4	2.4	25
2.5	2.5	26	2.5	2.5	26	2.5	2.5	26
2.6	2.6	27	2.6	2.6	27	2.6	2.6	27
2.7	2.7	28	2.7	2.7	28	2.7	2.7	28
2.8	2.8	29	2.8	2.8	29	2.8	2.8	29
2.9	2.9	30	2.9	2.9	30	2.9	2.9	30
3.0	3.0	31	3.0	3.0	31	3.0	3.0	31
3.1	3.1	32	3.1	3.1	32	3.1	3.1	32
3.2	3.2	33	3.2	3.2	33	3.2	3.2	33
3.3	3.3	34	3.3	3.3	34	3.3	3.3	34
3.4	3.4	35	3.4	3.4	35	3.4	3.4	35
3.5	3.5	36	3.5	3.5	36	3.5	3.5	36
3.6	3.6	37	3.6	3.6	37	3.6	3.6	37
3.7	3.7	38	3.7	3.7	38	3.7	3.7	38
3.8	3.8	39	3.8	3.8	39	3.8	3.8	39
3.9	3.9	40	3.9	3.9	40	3.9	3.9	40
4.0	4.0	41	4.0	4.0	41	4.0	4.0	41
4.1	4.1	42	4.1	4.1	42	4.1	4.1	42
4.2	4.2	43	4.2	4.2	43	4.2	4.2	43
4.3	4.3	44	4.3	4.3	44	4.3	4.3	44
4.4	4.4	45	4.4	4.4	45	4.4	4.4	45
4.5	4.5	46	4.5	4.5	46	4.5	4.5	46
4.6	4.6	47	4.6	4.6	47	4.6	4.6	47
4.7	4.7	48	4.7	4.7	48	4.7	4.7	48
4.8	4.8	49	4.8	4.8	49	4.8	4.8	49
4.9	4.9	50	4.9	4.9	50	4.9	4.9	50
5.0	5.0	51	5.0	5.0	51	5.0	5.0	51
5.1	5.1	52	5.1	5.1	52	5.1	5.1	52
5.2	5.2	53	5.2	5.2	53	5.2	5.2	53
5.3	5.3	54	5.3	5.3	54	5.3	5.3	54
5.4	5.4	55	5.4	5.4	55	5.4	5.4	55
5.5	5.5	56	5.5	5.5	56	5.5	5.5	56

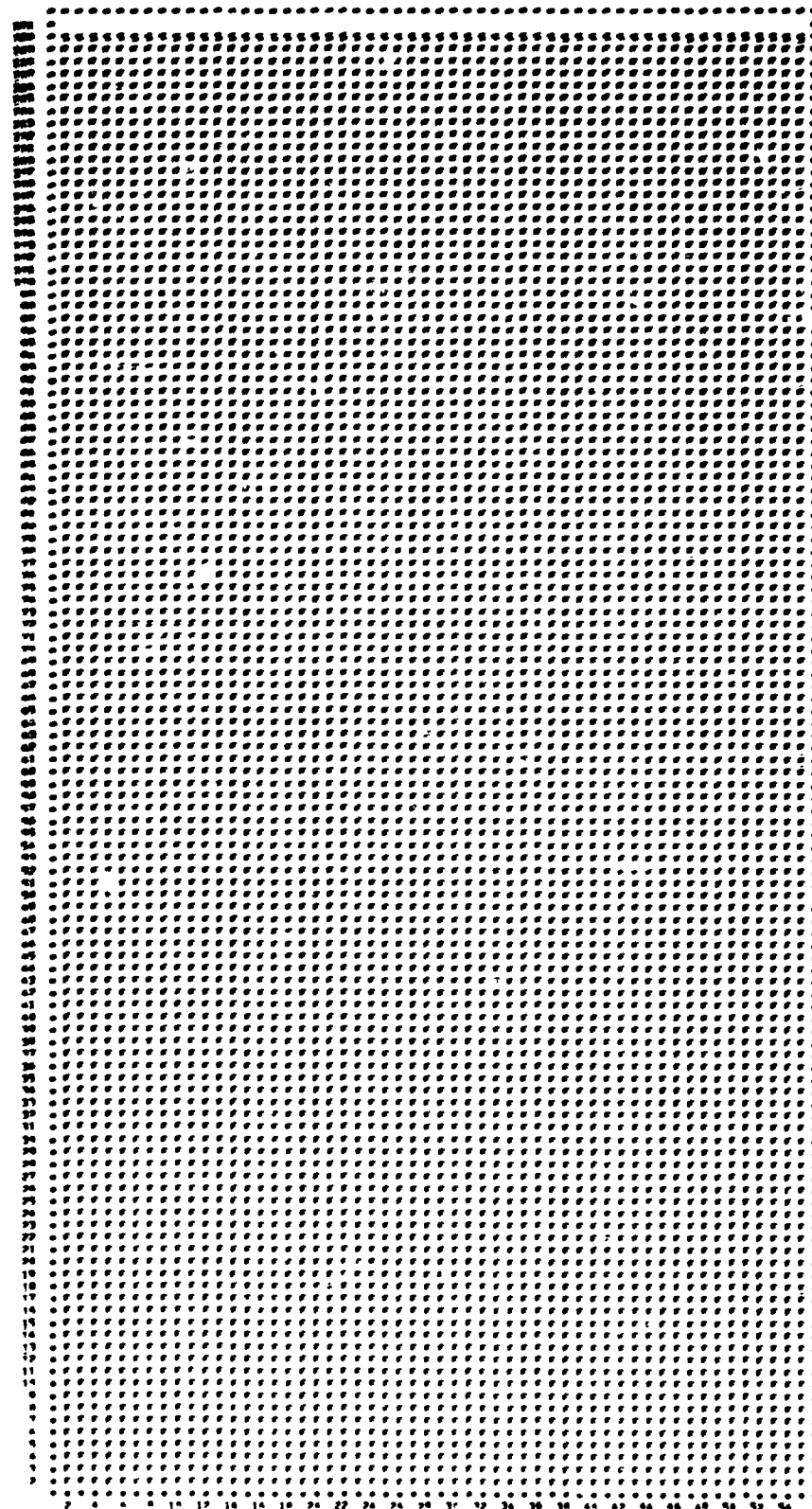


FIG. 4 CELL FLAGGING FOR PROBLEMS 1 AND 2

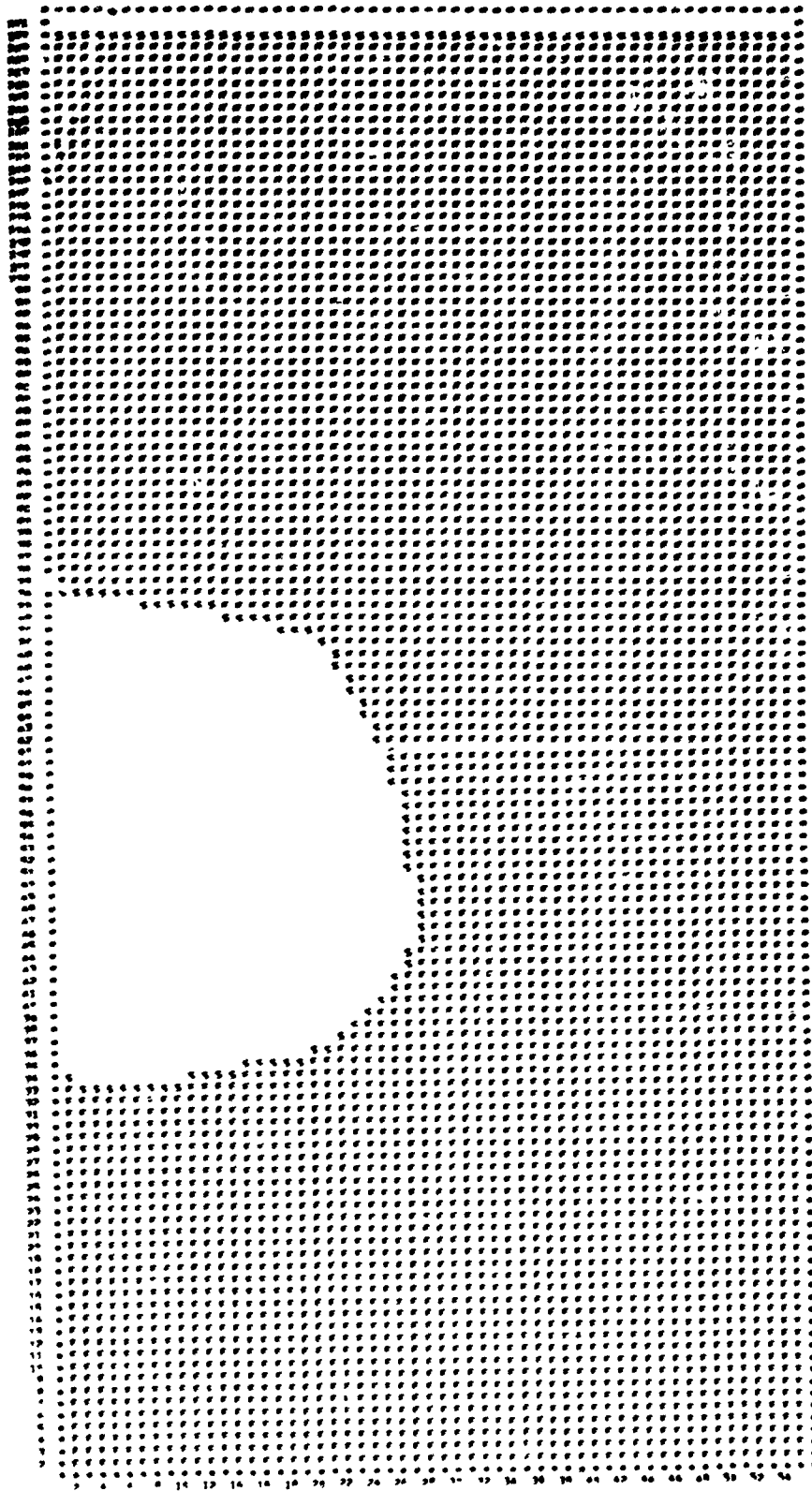


FIG. 5 CELL FLAGGING FOR PROBLEMS 3 AND 4

as an error estimate. The right hand side of (16) is not an error bound since for any matrix norm when  $\|L_u\| < 1$  we have

$$\frac{\rho(L_u)}{1 - \rho(L_u)} \leq \frac{\|L_u\|}{1 - \|L_u\|}$$

For all problems we have the following data:

$M$  = radial dimension of mesh = 56

$N$  = vertical dimension of mesh = 122

$J_s$  = index of horizontal layer of surface cells = 120

$\phi_{1,J_s} = 0$

$R_{1,j} = 0$

$w_j = 0$

$\phi_{1,j}$  = known true solution =  $-g(z_{J_s} - z_j)$

$\phi_{1,j}^0$  = initial guess = 0

Table 2 gives the associated parameters needed to calculate the error estimates for Problems 1,2,3 and 4. Tables 3 and 4 give Albrecht's two error estimates from equations (10) or (14), and (11) or (15), Tables 3 and 4 also give the spectral error estimate from equation (16) as well as the true error  $\|\phi_n - \phi\|_2$ . For comparison all parameters are normalized by  $\|\phi_n\|_2$ .

## CONCLUSION

From Tables 3 and 4 we see that for the unevenly spaced problems, Albrecht's bounds are over-estimates by more than five significant digits, and that for the evenly spaced problems they are over-estimates by about three significant digits. This is due to the fact that, as shown in Table 2,  $\|n^{-1/2}\|_2$  is large for the unevenly spaced problems. However, even for the evenly spaced problems, Albrecht's error bounds are overly pessimistic and thus computationally unusable.

Parameter	Problem			
	1	2	3	4
$\rho(B)$ (1)	.999949	.999980	.999949	.999864
$\omega$ (2)	1.980	1.987	1.980	1.968
$\rho(L_\omega)$ (3)	.980	.987	.980	.968
$\lambda$ (4)	9.80 E+3	2.53 E+4	9.80 E+3	3.66 E+3
$\ D^{-1/2}\ _2$ (5)	184.	.354	184.	.354
$\ D^{1/2}\ _2$ (5)	4.00	4.00	3.16	4.00
$\alpha$ (6)	1.80 E+6	8.96 E+3	1.80 E+6	5.20 E+3
$\beta$ (7)	7.21 E+6	3.58 E+4	5.70 E+6	5.20 E+3
$\delta$ (8)	49.0	75.9	49.0	30.3

- (1) Spectral radius of associated Jacobi Matrix
- (2) Optimum over-relaxation factor
- (3) Spectral radius of over-relaxation matrix
- (4) Constant from (10) and (11) for Albrecht's error estimate
- (5) D is the diagonal of  $L_\omega$
- (6) Constant from (10) and (11) where  $\alpha = \lambda \|D^{-1/2}\|_2$
- (7) Constant from (10) and (11) where  $\beta = \alpha \|D^{1/2}\|_2$
- (8)  $\delta = \rho(L_\omega)/(1-\rho(L_\omega))$  for spectral error estimate (16)

Table 2 Parameters Needed to Calculate Error Estimates

n	$  e_n  _2^{(1)}$	$  \phi_n - \phi  _2 /   \phi_n  _2$	$\beta   e_n  _2^{(2)}$	$\alpha   D^{1/2} e_n  _2^{(3)}$	$\delta   e_n  _2^{(4)}$
50	4.70 E-2	1.38	3.39 E+5	2.01 E+5	2.30
100	6.29 E-3	1.37 E-1	4.54 E+4	2.92 E+4	3.08 E-1
150	2.44 E-3	2.98 E-2	1.76 E+4	1.04 E+4	1.20 E-1
200	8.20 E-4	1.92 E-2	5.92 E+3	2.93 E+3	4.02 E-2
250	2.20 E-4	6.80 E-3	1.58 E+3	9.68 E+2	1.08 E-2
300	1.38 E-4	4.35 E-3	9.99 E+2	5.47 E+2	6.78 E-3
350	5.17 E-5	9.88 E-4	3.73 E+2	2.53 E+2	2.53 E-3
400	2.03 E-5	1.66 E-4	1.46 E+2	9.24 E+1	9.94 E-4

Problem 1 -- Unevenly Spaced Rectangular Mesh

n	$  e_n  _2^{(1)}$	$  \phi_n - \phi  _2 /   \phi_n  _2$	$\beta   e_n  _2^{(2)}$	$\alpha   D^{1/2} e_n  _2^{(3)}$	$\delta   e_n  _2^{(4)}$
50	2.63 E-2	1.78	9.62 E+2	9.59 E+2	2.12
100	7.87 E-3	3.58 E-1	2.82 E+2	2.81 E+2	6.22 E-1
150	3.27 E-3	4.69 E-2	1.17 E+2	1.16 E+2	2.53 E-1
200	5.93 E-4	5.45 E-2	2.12 E+1	2.11 E+1	4.68 E-2
250	4.27 E-4	4.66 E-2	1.53 E+1	1.52 E+1	3.37 E-2
300	4.77 E-4	2.63 E-2	1.71 E+1	1.70 E+1	3.76 E-2
350	3.36 E-4	6.63 E-3	1.21 E+1	1.20 E+1	2.66 E-2
400	8.75 E-5	2.22 E-3	3.14	3.12	6.91 E-3

Problem 2 -- Evenly Spaced Rectangular Mesh

- (1) To normalize results let  $e_n = (\phi_n - \phi_{n-1}) / ||\phi_n||_2$
- (2) Albrecht's error estimates from equations (11) and (15)
- (3) Albrecht's error estimates from equations (10) and (14)
- (4) Spectral error from equation (16) letting  $\delta = \rho(L_w) / (1 - \rho(L_w))$

Table 3 Errors for Problems 1 and 2

n	$\ e_n\ _2^{(1)}$	$\ \phi_n - \phi\ _2 / \ \phi_n\ _2$	$\beta \ e_n\ _2^{(2)}$	$\alpha \ D^{1/2} e_n\ _2^{(3)}$	$\delta \ e_n\ _2^{(4)}$
50	1.58 E-2	3.80 E-1	8.20 E+1	8.14 E+1	4.70 E-1
100	3.83 E-3	2.42 E-2	1.99 E+1	1.98 E+1	1.14 E-1
150	5.94 E-4	1.40 E-2	3.09	3.06	1.77 E-2
200	1.22 E-4	1.24 E-3	6.33 E-1	6.29 E-1	3.63 E-3
250	4.38 E-5	4.40 E-4	2.28 E-1	2.26 E-1	1.30 E-3
300	5.56 E-6	3.57 E-5	2.89 E-2	2.86 E-2	1.65 E-4
350	1.15 E-6	1.90 E-5	5.97 E-3	5.85 E-3	4.42 E-5
400	2.11 E-7	1.76 E-6	1.09 E-3	1.09 E-3	6.28 E-6

## Problem 4 -- Evenly Spaced Non-rectangular Mesh

n	$\ e_n\ _2^{(1)}$	$\ \phi_n - \phi\ _2 / \ \phi_n\ _2$	$\beta \ e_n\ _2^{(2)}$	$\alpha \ D^{1/2} e_n\ _2^{(3)}$	$\delta \ e_n\ _2^{(4)}$
50	4.36 E-2	1.04	2.49 E+5	1.90 E+5	2.14
100	5.83 E-3	9.40 E-2	3.32 E+4	2.68 E+4	2.85 E-1
150	2.41 E-3	1.53 E-2	1.37 E+4	1.02 E+4	1.18 E-1
200	9.03 E-4	1.78 E-2	5.16 E+3	3.59 E+3	4.43 E-2
250	2.79 E-4	5.55 E-3	1.59 E+3	1.32 E+3	1.37 E-2
300	1.51 E-4	3.53 E-3	8.65 E+2	6.38 E+2	7.43 E-3
350	4.99 E-5	5.20 E-4	2.85 E+2	2.42 E+2	2.45 E-3
400	2.15 E-5	1.26 E-4	1.23 E+2	9.89 E+1	1.05 E-3

## Problem 3 -- Unevenly Spaced Non-rectangular Mesh

- (1) To normalize results let  $e_n = (\phi_n - \phi_{n-1}) / \|\phi_n\|_2$
- (2) Albrecht's error estimates from equations (11) and (15)
- (3) Albrecht's error estimates from equations (10) and (14)
- (4) Spectral error from equation (16) letting  $\delta = \rho(L_\omega) / (1 - \rho(L_\omega))$

Table 4 Errors for Problems 3 and 4

A surprising computational result is that the spectral estimate is very good for all of the sample problems. We have hence decided to implement this much more practical but less desirable spectral error estimate (16) into the MACNOL computer program.

#### ACKNOWLEDGEMENTS

I would like to thank Dr. James Vander Graft of the University of Maryland and Dr. Andrew Van Tuyl of the Naval Ordnance Laboratory for reading the drafts of this report and for their invaluable suggestions.



**REFERENCES**

**MR** refers to Mathematical Reviews

- (1) **Albrecht, J.**, Fehlerabschaetzungen bei Relaxationsverfahren zur Numerischen Aufloesung Linearer Gleichungssysteme, Numer. Math. 3, 188-201 (1961).  
MR 26 No. 4476.
- (2) ———, Monotone Iterationsfolgen und ihre Verwendung zur Loesung linearer Gleichungssysteme, Numer. Math. 3, 345-358 (1961).
- (3) ———, Fehlerschranken und Konvergenzbeschleunigung bei einer monotonen oder alternierenden Iterationsfolge, Numer. Math. 4, 196-208 (1962).
- (4) ———, Zur Fehlerabschaetzung beim Gesamt- und Einzelschrittverfahren fur lineare Gleichungssysteme, Z. Angew. Math. 43, 83-95 (1963).  
MR 26 No. 5728.
- (5) **Collatz, L.**, Fehlerabschaetzung fur das Iterationsverfahren zur Aufloesung linearer Gleichungssysteme, Z. Angew. Math. Mech. 22, 357-361 (1942).
- (6) **Dueck, W.**, Eine Fehlerabschaetzung zum Einzelschrittverfahren bei linearen Gleichungssystemen, Numer. Math. 1, 73-77 (1959). MR 21 No. 1695.
- (7) **Feldman, H.**, Ein Hinreichendes Konvergenzkriterium und eine Fehlerabschaetzung fuer die Iteration in Einzelschritten bei linearen Gleichungssystemen, Z. Angew. Math. Mech. 48, 515-516 (1961). MR 25 No. 1638.
- (8) **Fitzgerald, K. E.**, Error Estimates for the Solution of Linear Algebraic Systems, J. Res. Nat. Bur. Stand., 74B (Math Sc.), No. 4, 251-260 (Oct. - Dec. 1970).
- (9) **Householder, A.**, "The Theory of Matrices in Numerical Analysis," Blaisdell Publishing Company, New York, 1964.
- (10) **Pritchett, J.**, MACYL--A Two Dimension Cylindrical Coordinate Incompressible Hydrodynamic Code, U.S. Naval Radiological Defense Laboratory, USNRDL-TR-67-97 (1967).
- (11) **Sassenfeld, H.**, Ein hinreichendes Konvergenzkriterium und eine Fehlerabschaetzung fur die Iteration in Einzelschritten bei linearen Gleichungen, Z. Angew. Math. Mech. 31, 92-94 (1951). MR 14 p. 692.

- (12) Schmidt, J. W., Fehlerabschaetzungen und Konvergenzbeschleunigung zu iterationen bei linearen Gleichungssystemen, Apl. Mat. 10, 297-301 (1965). MR 32 No. 8488.
- (13) Schroeder, J., Neue Fehlerabschaetzungen fur verschiedene Iterationsverfahren, Z. Angew. Math. Mech. 36, 168-181 (1956). MR 18 p. 152.
- (14) ———, Fehlerabschaetzung bei linearen Gleichungssystemen mit dem BROUWERSchen Fixpunktsatz, Arch. Rat. Mech. Anal. 3, 28-44 (1959).
- (15) ———, Computing Error Bounds in Solving Linear Systems, Math. Comput. 16, 323-337 (1962). MR 26 No. 7147.
- (16) Varga, R. S., "Matrix Iterative Analysis," Prentice-Hall, Englewood Cliffs, New Jersey, 1962.
- (17) von Mises, R. V. and Pollaczek-Geiringer, H., Praktische Verfahren der Gleichungsaufloesung, Z. Angew. Math. Mech. 9, 58-73 (1929).
- (18) Weinberger, H.F., A Posteriori Error Bounds in Iterative Matrix Inversion, Numerical Solution of Partial Differential Equations, Proc. Sympos. Univ. Maryland, 153-163 (1966). MR 34 No. 8362.
- (19) Weissinger, J., Zur Theorie und Anwendung des Iterationsverfahren, Z. Angew. Math. Mech. 36, 168-181 (1956).
- (20) Young, D., Iterative Methods for Solving Partial Difference Equations of Elliptic Type, Trans. Amer. Math. Soc. 76, No.1, 92-111 (1954).
- (21) ———, "Iterative Solution of Large Linear Systems," Academic Press, New York and London, 1971.